Adaptative Hausdorff Distances and Dynamic Clustering of Symbolic Interval Data

Francisco de A.T. de Carvalho, Renata M.C.R. de Souza, Marie Chavent, Yves Lechevallier

Adaptive Hausdorff Distances and Dynamic Clustering of Symbolic Interval Data

Francisco de A.T. de Carvalho [a,*], Renata M.C.R. de Souza [a], Marie Chavent [b], Yves Lechevallier [c]

[a] *Centro de Informática, Universidade Federal de Pernambuco, Caixa Postal 7851 - CEP 50732-970 - Recife (PE) - Brazil*

[b] *Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS, Université Bordeaux 1 - 351, Cours de la liberation, 33405 Talence Cedex, France*

[c] *INRIA-Institut National de Recherche en Informatique et en Automatique Domaine de Voluceau-Rocquencourt B. P.105, 78153 Le Chesnay Cedex, France*

Abstract

This paper presents a partitional dynamic clustering method for interval data based on adaptive Hausdorff distances. Dynamic clustering algorithms are iterative two-step relocation algorithms involving the construction of the clusters at each iteration and the identification of a suitable representation or prototype (means, axes, probability laws, groups of elements, etc.) for each cluster by locally optimizing an adequacy criterion that measures the fitting between the clusters and their corresponding representatives. In this paper, each pattern is represented by a vector of intervals. Adaptive Hausdorff distances are the measures used to compare two interval vectors. Adaptive distances at each iteration change for each cluster according to its intra-class structure. The advantage of these adaptive distances is that the clustering algorithm is able to recognize clusters of different shapes and sizes. To evaluate this method, experiments with real and synthetic interval data sets were performed. The evaluation is based on an external cluster validity index (corrected Rand index) in a framework of a Monte Carlo experiment with 100 replications. These experiments showed the usefulness of the proposed method.

Key words: Symbolic data analysis, dynamic clustering, interval data, Hausdorff distance, adaptive distances.

* Corresponding Author. tel.:+55-81-21268430; fax:+55-81-21268438

Email addresses: fatc@cin.ufpe.br (Francisco de A.T. de Carvalho), rmcrs@cin.ufpe.br (Renata M.C.R. de Souza), Marie.Chavent@math.u-bordeaux.fr (Marie Chavent), Yves.Lechevallier@inria.fr (Yves Lechevallier).

[1] Acknowledgements. The first two authors would like to thank CNPq (Brazilian

1 Introduction

Cluster analysis seeks to organize a set of items (usually represented as a vector of quantitative values in a multidimensional space) into clusters such that items within a given cluster have a high degree of similarity, whereas items belonging to different clusters have a high degree of dissimilarity (Bock (1993), Jain et al (1999)). Cluster analysis techniques can be divided into hierarchical and partitional methods (Spaeth (1980), Gordon (1999), Everitt (2001)).

Hierarchical methods yield complete hierarchy, i.e., a nested sequence of partitions of the input data. Hierarchical methods can be agglomerative or divisive. Agglomerative methods yield a sequence of nested partitions starting with trivial clustering in which each item is in a unique cluster and ending with trivial clustering in which all items are in the same cluster. A divisive method starts with all items in a single cluster and performs a splitting procedure until a stopping criterion is met (usually upon obtaining a partition of singleton clusters).

Partitional methods seek to obtain a single partition of the input data into a fixed number of clusters. They usually produce clusters by (locally) optimizing an adequacy criterion. To improve cluster quality, the algorithm is run multiple times with different starting points and the best configuration obtained from the total runs is used as the output clustering.

This paper addresses the partitioning of interval data often present in real applications. Table 1 shows an example of an interval data table.

Table 1
Cardiological interval data set

u	Pulse rate	Systolic blood pressure	Diastolic blood pressure
1	[44-68]	[90-100	[50-70]
2	[60-72]	[90-130]	[70-90]
...	...	...	...
10	[86-96]	[138-180]	[90-110]
11	[86-100]	[110-150]	[78-100]

This kind of data have been studied mainly in *Symbolic Data Analysis* (SDA), a new domain in the area of knowledge discovery and data management related to multivariate analysis, pattern recognition and artificial intelligence. The aim of SDA is to provide suitable methods (clustering, factorial techniques,

Agency) for its financial support.

decision trees, etc.) for managing aggregated data described by multi-valued variables, where the cells of the data table contain sets of categories, intervals, or weight (probability) distributions (Diday (1988), Bock and Diday (2000), Billard and Diday (2003)).

SDA has provided partitioning methods for clustering symbolic data. Diday and Brito (1989) proposed a clustering approach based on a transfer algorithm. El-Sonbaty and Ismail (1998) proposed a fuzzy k-means algorithm for clustering different types of symbolic data. Verde et al (2001) introduced a dynamic cluster algorithm for symbolic data considering context-dependent proximity functions. Gordon, A. D. (2000) presented an iterative relocation algorithm that minimizes the sum of the description potentials of the clusters. Bock (2002) gives clustering strategies based on a clustering criterion and presents a sequential clustering and updating strategy for constructing a Self-Organizing Map in order to visualize symbolic interval-type data. Chavent and Lechevallier (2002) proposed a dynamical clustering algorithm for interval data where the prototype is defined by the optimization of an adequacy criterion based on the Hausdorff distance. Moreover, in Souza and De Carvalho (2004), an adaptive dynamic clustering algorithm is presented for interval data based on City-block distances.

The main contribution of this paper is the proposal of a new partitional dynamic clustering method for interval data based on the use of an adaptive Hausdorff distance at each iteration.

The partitioning dynamical cluster algorithms (Diday (1971)) are iterative two-step relocation algorithms involving the construction of clusters at each iteration and the identification of a suitable representation or prototype (means, axes, probability laws, groups of elements, etc.) for each cluster by locally optimizing an adequacy criterion between the clusters and their corresponding representations (Diday and Simon (1976)). An allocation step is performed to assign individuals to classes according to their proximity to the prototypes. This is followed by a representation step where the prototypes are updated according to the assignment of the individuals in the allocation step, until the convergence of the algorithm, when the adequacy criterion reaches a stationary value.

The idea of dynamical clustering with adaptive distances (Govaert (1975), Diday and Govaert (1977)) is to associate a distance to each cluster, which is defined according to its intra-class structure. The advantage of this approach is that the clustering algorithm recognizes different shapes and sizes of clusters. In this paper, the adaptive distance is a weighted sum of Hausdorff distances. Explicit formulas for the optimum class prototype, as well as for the the weights of the adaptive distances, are found. When used for dynamic clustering of interval data, these prototypes and weights ensure that the clustering

criterion decreases at each iteration.

In this paper, we present a dynamic clustering method with adaptive Hausdorff distances for partitioning a set of interval data. This method is an extension of the dynamic clustering algorithm based on non-adaptive Hausdorff distances proposed in Chavent and Lechevallier (2002). In Section 2, a description is given of the classical dynamic clustering method with adaptive distances. This is followed by the presentation of the dynamic clustering method based on adaptive Hausdorf distances for interval data (Section 3). To validate this new method, Section 4 presents experiments with real and synthetic interval data sets. In Section 5, the concluding remarks are given.

2 Introduction to partitional dynamic clustering with adaptive distances

Let Ω be a set n of objects indexed by i and described by p variables indexed by j. Each object i is represented by a vector of feature values $\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$. Throughout this paper, we consider the problem of clustering Ω into K disjoint clusters $C_1, ..., C_K$ such that the resulting partition $P = (C_1, ..., C_K)$ is optimum with respect to a given clustering criteria.

In dynamic clustering (Diday and Simon (1976)), we represent each cluster $C_k \in P$ by a prototype $\mathbf{y}_k$, which is also a vector of feature values. We measure the quality of this cluster by the sum of the dissimilarities $d(\mathbf{x}_i, \mathbf{y}_k)$ between objects $i \in C_k$ and the prototype $\mathbf{y}_k$. This measure of quality $\sum_{i \in C_k} d(\mathbf{x}_i, \mathbf{y}_k)$ is called the adequacy criterion of the cluster C_k. The classification problem is to find a partition P and a set L of K prototypes that minimize the following clustering criterion:

$$\Delta(P, L) = \sum_{i=1}^{K} \sum_{i \in C_K} d(\mathbf{x}_i, \mathbf{y}_k) \tag{1}$$

over all partitions $P = (C_1, ..., C_K)$ of Ω and all choices of set $L = (\mathbf{y}_1, ...\mathbf{y}_K)$ of cluster prototypes.

In this context, the dynamic clustering algorithm iteratively performs both a *representation step* and an *allocation step*:

a) Representation step (the partition P is fixed).

Finding L that minimizes $\Delta(P, \bullet)$ is equivalent to finding for $k = 1, ..., K$, the prototype $\mathbf{y}_k$ that minimizes the adequacy criterion $\sum_{i \in C_k} d(\mathbf{x}_i, \mathbf{y}_k)$. For con-

tinuous quantitative data in $\Re^p$ and the city-block distance, $\mathbf{y}_k$ is the median vector of the cluster C_k.

b) Allocation step (the set of prototypes L is fixed).

Finding P that minimizes $\Delta(\bullet, L)$ is equivalent to finding for $k = 1, ..., K$, the cluster $C_k = \{i \in \Omega \mid d(\mathbf{x}_i, \mathbf{y}_k) \leq d(\mathbf{x}_i, \mathbf{y}_m)\ , \forall m = 1, ..., K\}$

Once these two steps properly defined, the partitioning criterion (1) decreases at each iteration and the algorithm converges to a stationary value of this criterion under the two following conditions:

i) Unicity of the choice for the cluster affectation of each object of Ω;
ii) Unicity of the choice of the prototype $\mathbf{y}_k$ that minimizes the adequacy criterion $\sum_{i \in C_k} d(\mathbf{x}_i, \mathbf{y}_k)$

The main idea of dynamic clustering with adaptive distances is to associate a distance d_k to each cluster C_k (and its prototype $\mathbf{y}_k$) such that the sum of the distances $d_k(\mathbf{x}_i, \mathbf{y}_k)$ between objects $i \in C_k$ and the prototype $\mathbf{y}_k$ is as small as possible. The distances used in the dynamic algorithm are therefore not determined once and for all. Moreover, they are different from one cluster to another. The clustering criterion is:

$$\Delta(P, L) = \sum_{k=1}^{K} \sum_{i \in C_k} d_k(\mathbf{x}_i, \mathbf{y}_k) \tag{2}$$

where $P = (C_1, ..., C_K)$ and now $L = (G, d)$, where $G = (\mathbf{y}_1, ..., \mathbf{y}_K)$ and $d = (d_1, \ldots, d_K)$.

In our context, the distance d_k is a weighted sum of distances d^j, where d^j compares a pair of objects according to variable j:

$$d_k(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} d^j(x_i^j, x_{i'}^j) = \sum_{j=1}^{p} \lambda_k^j\ d(x_i^j, x_{i'}^j) \tag{3}$$

with $d^j(x_i^j, x_{i'}^j) = \lambda_k^j\ d(x_i^j, x_{i'}^j)$, $\lambda_k^j > 0$ and $\prod_{j=1}^{p} \lambda_k^j = 1$.

According to the definition of d_k given in (3), the set L is written $L = (G, \lambda)$ where $\lambda = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$, with $\boldsymbol{\lambda}_k = (\lambda_i^1, ..., \lambda_i^p)$ being the vector of weights of the fixed distance d. The adaptivity of the distance d_k is expressed by the vector of weights $\boldsymbol{\lambda}_k$.

When using adaptive distances, the *representation step* is divided in two stages:

a1) Stage 1 (the partition P and λ are fixed).

Find for $k = 1, ..., K$, the prototype $\mathbf{y}_k$ that minimizes the adequacy criterion $\sum_{i \in C_k} d_k(\mathbf{x}_i, \mathbf{y}_k) = \sum_{i \in C_k} \sum_{j=1}^{p} \lambda_k^j \, d(x_i^j, y_k^j) = \sum_{j=1}^{p} \lambda_k^j \, \sum_{i \in C_k} d(x_i^j, y_k^j)$.

a2) Stage 2 (the partition P and the set of prototypes G are fixed).

Find for $k = 1, ..., K$, the vector of weights $\boldsymbol{\lambda}_k$ that minimizes the adequacy criterion $\sum_{i \in C_k} d_k(\mathbf{x}_i, \mathbf{y}_k) = \sum_{j=1}^{p} \lambda_k^j \sum_{i \in C_k} d(x_i^j, y_k^j) = \sum_{j=1}^{p} \lambda_k^j \Phi_j$ where $\Phi_j = \sum_{i \in C_k} d(x_i^j, y_k^j)$.

The *allocation step* of the algorithm is once again:

b) Allocation step (the set of prototypes G and and the vector λ are fixed):

Find for $k = 1, ..., K$, $C_k = \{i \in \Omega \mid d_k(\mathbf{x}_i, \mathbf{y}_k) \leq d_m(\mathbf{x}_i, \mathbf{y}_m) \, , \forall m = 1, ..., K\}$.

Once these two steps have been properly defined, the partitioning criterion (2) decreases at each iteration and the algorithm converges to a stationary value of this criterion under the three following conditions:

i) Unicity of the choice for the cluster affectation of each object of Ω;
ii) Unicity of the choice of the prototype $\mathbf{y}_k$ that minimizes the adequacy criterion $\sum_{i \in C_k} d_k(\mathbf{x}_i, \mathbf{y}_k)$
iii) Unicity of the choice of the vectors of weights $\boldsymbol{\lambda}_k$ that minimizes the adequacy criterion $\sum_{j=1}^{p} \lambda_k^j \sum_{i \in C_k} d(x_i^j, y_k^j)$.

3 Dynamic clustering method with an adaptive Hausdorff distances

In this paper we are concerned with objects that are represented by a vector of intervals (we consider a point as an interval with equal lower and upper bounds). Let Ω be a set of n objects indexed by i and described by p interval variables indexed by j. An *interval variable* X (Bock and Diday (2000)) is a correspondence defined from Ω in $\Re$ such that for each $i \in \Omega, X(i) = [a, b] \in \Im$, where $\Im$ is the set of closed intervals defined from $\Re$.

Each object i is represented as a vector of intervals $\mathbf{x}_i = (x_i^1, \cdots, x_i^p)$, where $x_i^j = [a_i^j, b_i^j] \in \Im = \{[a, b] : a, b \in \Re, \ a \leq b\}$. In this paper, an interval data table $\{x_i^j\}_{n \times p}$ is made up of n rows that represent n objects to be clustered and p columns that represent p interval variables. Each cell of this table contains an interval $x_i^j = [a_i^j, b_i^j] \in \Im$.

A prototype $\mathbf{y}_k$ of cluster $C_k \in P$ is also represented as a vector of intervals $\mathbf{y}_k = (y_k^1, \cdots, y_k^p)$, where $y_k^j = [\alpha_k^j, \beta_k^j] \in \Im$.

It is now a matter of choosing an adaptive distance between vectors of intervals and properly defining the representation step of the dynamic algorithm with adaptive distances given in the previous section. In other words, we will give an explicit formula for the prototype $\mathbf{y}_k$ and for the vector of weights $\boldsymbol{\lambda}_k$ that minimizes both the adequacy criterion $\sum_{j=1}^{p} \lambda_k^j \sum_{i \in C_k} d(x_i^j, y_k^j)$.

3.1 The adaptive Hausdorff distances for interval data

A number of proximity measures have been introduced in the literature for interval data (as well as for other types of symbolic data). Gowda and Diday (1991) and Gowda and Diday (1992) introduced, respectively, dissimilarity and similarity functions with components based on position, span and content. The component based on position indicates the relative positions of two feature values on a real line. The component based on span indicates the relative sizes of the feature values without referring their common parts. The component based on content is a measure of the common parts between two features values. Ichino and Yaguchi (1994) presented the generalized Minkowski metrics for mixed feature variables. Similarity and dissimilarity measures between symbolic data restricted by dependency rules between feature values can be found in De Carvalho (1994), De Carvalho (1998) and De Carvalho and Souza (1998).

We have seen that the distance d_k associated with the cluster C_k is defined as a weighted sum of distances d^j, where d^j compares a pair of objects according to variable j:

$$d^j(x_i^j, x_{i'}^j) = \lambda_k^j \; d(x_i^j, x_{i'}^j)$$

Here, the two feature values x_i^j and $x_{i'}^j$ are, respectively, the two intervals $[a_i^j, b_i^j]$ and $[a_{i'}^j, b_{i'}^j]$. The distance d (see equation 3) chosen to compare two intervals is the Hausdorff distance. The Hausdorff distance (Nadler (1978), Rote (1991)) is often used in image processing (Huttenlocher et al. (1993)) and is defined to compare two sets of objects A and B. This distance depends on the distance chosen to compare two objects u and v respectively in A and B. We consider the euclidean distance and the Hausdorff distance is defined by:

$$d_H(A, B) = \max(h(A, B), h(B, A))$$

where:

$$h(A, B) = \sup_{u \in A} \inf_{v \in B} ||u - v||$$

At times, h is called the *directed* Hausdorff *distance*.

In this work, A and B are two intervals $x_i^j = [a_i^j, b_i^j]$ and $x_{i'}^j = [a_{i'}^j, b_{i'}^j]$. The previous Hausdorff distance is simplified to:

$$d_H(x_i^j, x_{i'}^j) = max\{|a_i^j - a_{i'}^j|, |b_i^j - b_{i'}^j|\} \tag{4}$$

Finally, the adaptive distance d_k associated with the cluster C_k and defined in equation (3) is:

$$d_k(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} \lambda_k^j \, d_H(x_i^j, x_{i'}^j) = \sum_{j=1}^{p} \lambda_k^j \, max\{|a_i^j - a_{i'}^j|, |b_i^j - b_{i'}^j|\} \tag{5}$$

with $\lambda_k^j > 0$ and $\prod_{j=1}^{p} \lambda_k^j = 1$.

3.2 Definition of the best prototypes

In section 2, we saw that the representation step of the dynamic clustering algorithm with adaptive distances was divided into two stages, corresponding to two minimization problems. The first problem, when the partition P and the vector λ are fixed, is to find for $k = 1, ..., K$ the prototype $\mathbf{y}_k$ that minimizes the adequacy criterion $\sum_{i \in C_k} d_k(\mathbf{x}_i, \mathbf{y}_k)$. With d_k defined in (5) and with $\mathbf{x}_i = (x_i^1, \cdots, x_i^p)$ and $\mathbf{y}_k = (y_k^1, \cdots, y_k^p)$, the adequacy criterion is:

$$\begin{aligned} \sum_{i \in C_k} d_k(\mathbf{x}_i, \mathbf{y}_k) &= \sum_{i \in C_k} \sum_{j=1}^{p} \lambda_k^j \, d_H(x_i^j, y_k^j) \\ &= \sum_{j=1}^{p} \lambda_k^j \sum_{i \in C_k} d_H(x_i^j, y_k^j) \end{aligned} \tag{6}$$

The vector of weights being fixed, the problem is now to find for $j = 1, \ldots, p$ the interval $y_k^j = [\alpha_k^j, \beta_k^j]$ that minimizes:

$$\sum_{i \in C_k} d_H(x_i^j, y_k^j) = \sum_{i \in C_k} max\{|a_i^j - \alpha_k^j|, |b_i^j - \beta_k^j|\} \tag{7}$$

According to Chavent and Lechevallier (2002), an explicit formula for the components y_k^j of the best prototype is found by transforming the previous minimization problem into two well-known L_1 norm problems.

Let $m_i^j = (a_i^j + b_i^j)/2$ be the midpoint of the interval $x_i^j = [a_i^j, b_i^j]$ (for $j = 1, \ldots, p$) and $l_i^j = (b_i^j - a_i^j)/2$ be half of its length. From this we have:

$$a_i^j = m_i^j - l_i^j \text{ and } b_i^j = m_i^j + l_i^j \tag{8}$$

Also, let $\mu_k^j = (\alpha_k^j + \beta_k^j)/2$ be the midpoint of the interval $y_k^j = [\alpha_k^j, \beta_k^j]$ (for $j = 1, \ldots, p$) and $\rho_k^j = (\beta_k^j - \alpha_k^j)/2$ be half of its length. We have:

$$\alpha_k^j = \mu_k^j - \rho_k^j \text{ and } \beta_k^j = \mu_k^j + \rho_k^j \tag{9}$$

From the equations (7), (8) and (9), the equation (6) can be written as:

$$\begin{aligned}\sum_{i \in C_k} d_H(x_i^j, y_k^j) &= \sum_{i \in C_k} max\{|(m_i^j - l_i^j) - (\mu_k^j - \rho_k^j)|, |(m_i^j + l_i^j) - (\mu_k^j + \rho_k^j)|\} \\ &= \sum_{i \in C_k} max\{|(m_i^j - \mu_k^j) - (l_i^j - \rho_k^j)|, |(m_i^j - \mu_k^j) + (l_i^j - \rho_k^j)|\}\end{aligned}$$

According to the following property defined for x and y in $\Re$,

$$max(|x - y|, |x + y|) = |x| + |y|$$

Then:

$$\begin{aligned}\sum_{i \in C_k} d_H(x_i^j, y_k^j) &= \sum_{i \in C_k} (|m_i^j - \mu_k^j| + |l_i^j - \rho_k^j|) \\ &= \sum_{i \in C_k} |m_i^j - \rho_k^j| + \sum_{i \in C_k} |l_i^j - \rho_k^j|\end{aligned} \tag{10}$$

This yields two well-known minimization problems in L_1 norm: find $\mu_k^j \in \Re$ and $\rho_k^j \in \Re$ that respectively minimize:

$$\sum_{i \in C_k} |m_i^j - \mu_k^j| \text{ and } \sum_{i \in C_k} |l_i^j - \rho_k^j|$$

The solution $\hat{\mu}_k^j$ and $\hat{\rho}_k^j$ are, respectively, the median of the set $\{m_i^j, i \in C_k\}$ (the midpoints of the intervals $x_i^j = [a_i^j, b_i^j], i \in C_k$), and the median of the set $\{l_i^j, i \in C_k\}$ (the half-lengths of the intervals $x_i^j = [a_i^j, b_i^j], i \in C_k$). Finally, the solution $\hat{y}_k^j = [\hat{\alpha}_k^j, \hat{\beta}_k^j]$ is given by

$$\hat{\alpha}_k^j = \hat{\mu}_k^j - \hat{\rho}_k^j \text{ and } \hat{\beta}_k^j = \hat{\mu}_k^j + \hat{\rho}_k^j \tag{11}$$

3.3 Definition of the best distances

The second stage of the representation step of the dynamic clustering algorithm with adaptive distances, when the partition P and the set of prototypes G are fixed, is to find for $k = 1, ..., K$ the vector of weights $\boldsymbol{\lambda}_k$ that minimizes the adequacy criterion defined in (6) by:

$$\sum_{i \in C_k} d_k(\mathbf{x}_i, \mathbf{y}_k) = \sum_{j=1}^{p} \lambda_k^j \Phi_j \quad \text{where} \quad \Phi_j = \sum_{i \in C_k} d_H(x_i^j, y_k^j)$$

Following Diday and Govaert (1974) and Govaert (1975), the weights λ_k^j are calculated by the Lagrange multiplier method:

$$\frac{\partial}{\partial \lambda_k^j}(\sum_{j=1}^{p} \lambda_k^j \Phi_j - \mu \prod_{h=1}^{p} \lambda_k^h) = 0 \text{ for } j = 1, \ldots, p \tag{12}$$

From equation (12), we have the following result:

$$\Phi_j - \mu \frac{\prod_{h=1}^{p} \lambda_k^h}{\lambda_k^j} = 0 \Rightarrow \lambda_k^j = \frac{\mu}{\Phi_j}(\prod_{h=1}^{p} \lambda_k^h) \tag{13}$$

Remembering that $\Pi_{h=1}^{p} \lambda_k^h = 1$, the parameter λ_k^j in equation (13) is given by

$$\lambda_k^j = \frac{\mu}{\Phi_j}$$

The restriction $\Pi_{h=1}^{p} \lambda_k^h = 1$ can be written as:

$$1 = \prod_{h=1}^{p} \frac{\mu}{\Phi_h} = \frac{\mu^p}{\prod_{h=1}^{p} \Phi_h} \text{ and then } \mu = (\prod_{h=1}^{p} \Phi_h)^{\frac{1}{p}}$$

Finally, the solution $\hat{\lambda}_k^j$ to the parameter λ_k^j is:

$$\hat{\lambda}_k^j = \frac{\mu}{\Phi_j} = \frac{\left[\prod_{h=1}^{p}(\sum_{i \in C_h} max\{|a_i^h - \hat{\alpha}_k^h|, |b_i^h - \hat{\beta}_k^h|\})\right]^{\frac{1}{p}}}{\sum_{i \in C_k} max\{|a_i^j - \hat{\alpha}_k^j|, |b_i^j - \hat{\beta}_k^j|\}} \tag{14}$$

3.4 The algorithm

The algorithm schema of dynamic clustering algorithm with Hausdorff adaptive distances for interval data is as follows:

Schema of adaptive dynamic clustering algorithm

(1) **Initialization**
Choose a partition $\{C_1 \ldots, C_K\}$ of Ω randomly or choose K distinct objects $\mathbf{y}_1, \ldots, \mathbf{y}_K$ belonging to Ω and assign each object i to the closest object $\mathbf{y}_{k*}$ ($k* = arg\ \min_{k=1,\ldots,K} \sum_{j=1}^{p} max\{|a_i^j - \alpha_k^j|, |b_i^j - \beta_k^j|\}$ to construct the initial partition $\{C_1, \ldots, C_K\}$.

(2) **Representation step**

a) For $i = 1$ to K compute the prototype $\hat{\mathbf{y}}_k = ([\hat{\alpha}_k^1, \hat{\beta}_k^1], \ldots, \hat{\alpha}_k^p, \hat{\beta}_k^p])$ with $\hat{\alpha}_k^j = \hat{\mu}_k^j - \hat{\rho}_k^j$ and $\hat{\beta}_k^j = \hat{\mu}_k^j + \hat{\rho}_k^j$ where $\hat{\mu}_k^j$ is the median of the set $\{m_i^j, i \in C_k\}$ and $\hat{\rho}_k^j$ is the median of the set $\{l_i^j, i \in C_k\}$

b) For $j = 1, \ldots, p$ and $k = 1, \ldots, K$, compute $\hat{\lambda}_k^j$ with equation (14)

(3) **Allocation step**

$test \leftarrow 0$
for $i = 1$ to n do
define the winning cluster C_{k*} such that

$$k* = arg\ \min_{k=1,\ldots,K} \sum_{j=1}^{p} \hat{\lambda}_k^j\, max\{|a_i^j - \hat{\alpha}_k^j|, |b_i^j - \hat{\beta}_k^j|\}$$

if $i \in C_k$ and $k* \neq k$
$test \leftarrow 1$
$C_{k*} \leftarrow C_{k*} \cup \{i\}$
$C_k \leftarrow C_k \setminus \{s\}$

(4) **Stopping criterion**
If $test = 0$ then STOP, otherwise go to (2).

For classical dynamical clustering methods, the initialization step and stopping rules can be modified. For example, the points chosen randomly at initialization can be chosen in such a way that they are as dissimilar as possible. Concerning the stopping rule, a minimum value for the clustering criterion or a maximum number of iterations can also be given.

Another remark is when all the weights λ_k^j are fixed to 1, the distances are non-adaptive and the previous algorithm is equivalent to the dynamic clustering algorithm of interval data proposed in Chavent and Lechevallier (2002). This remark will be used in the next section for evaluating the adaptive dynamic clustering algorithm.

4 Experimental results

To show the usefulness of this method, two synthetic interval data sets with linearly non-separable clusters of different shapes and sizes have been drawn.

Real applications are then considered.

Our aim is to achieve a comparison of the dynamic clustering algorithm considering different distances between vectors of intervals: adaptive Hausdorff distance proposed in this paper, non-adaptive Hausdorff distance (Chavent and Lechevallier (2002)), one component adaptive city-block distance (Souza and De Carvalho (2004)) and non-adaptive city-block distance.

To compare the results furnished by the dynamic clustering algorithm with these different distances, an external validity index is used. For synthetic interval data sets, rectangles are built from three clusters of points drawn from three bi-variate normal distributions. Next, the *a priori* partition of the objects is known. For the car interval data set describing car models, it is defined a *a priori* partition into four groups according to a *car category*. For the interval data set describing species of freshwater fish, it is considered a *a priori* partition of the species into four groups according to *diet*.

The idea of external validity is simply to compare the a priori partition with the partition obtained from the clustering algorithm. In this paper, we use the corrected Rand (CR) index defined in Hubert and Arabie (1985) for comparing two partitions, the definition of which is as follows.

Let $U = \{u_1, \ldots, u_i, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_j, \ldots, v_C\}$ be two partitions of the same data set having respectively R and C clusters. The corrected Rand index is:

$$CR = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \qquad (15)$$

where $\binom{n}{2} = \frac{n(n-1)}{2}$ and n_{ij} represents the number of objects that are in clusters u_i and v_i; $n_{i.}$ indicates the number of objects in cluster u_i; $n_{.j}$ indicates the number of objects in cluster v_j; and n is the total number of objects in the data set. CR takes its values from the interval [-1,1], where the value 1 indicates perfect agreement between partitions, whereas values near 0 (or negatives) correspond to cluster agreement found by chance.

4.1 Synthetic interval data sets

In this paper, we consider the same data point configurations presented in Souza and De Carvalho (2004). Two data sets of 350 points in $\Re^2$ were constructed. In each data set, the 350 points are drawn from three bi-variate normal distributions of independent components. There are three clusters of

unequal sizes and shapes: two clusters with an ellipsoidal shape and size 150 and one cluster with a spherical shape and size 50. The mean vector and the covariance matrix of the bi-variate normal distributions are noted:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ and } \Sigma_{11} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Data set 1 shows well-separated clusters (Figure 1). The data points of each cluster in this data set were drawn according to the following parameters:

a) Class 1: $\mu_1 = 28$, $\mu_2 = 22$, $\sigma_1^2 = 100$ and $\sigma_2^2 = 9$;
b) Class 2: $\mu_1 = 60$, $\mu_2 = 30$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 144$;
c) Class 3: $\mu_1 = 45$, $\mu_2 = 38$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 9$;

Data set 2 shows overlapping clusters (Figure 1). The data points of each cluster in this data set were drawn according to the following parameters:

a) Class 1: $\mu_1 = 45$, $\mu_2 = 22$, $\sigma_1^2 = 100$ and $\sigma_2^2 = 9$;
b) Class 2: $\mu_1 = 60$, $\mu_2 = 30$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 144$;
c) Class 3: $\mu_1 = 52$, $\mu_2 = 38$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 9$;

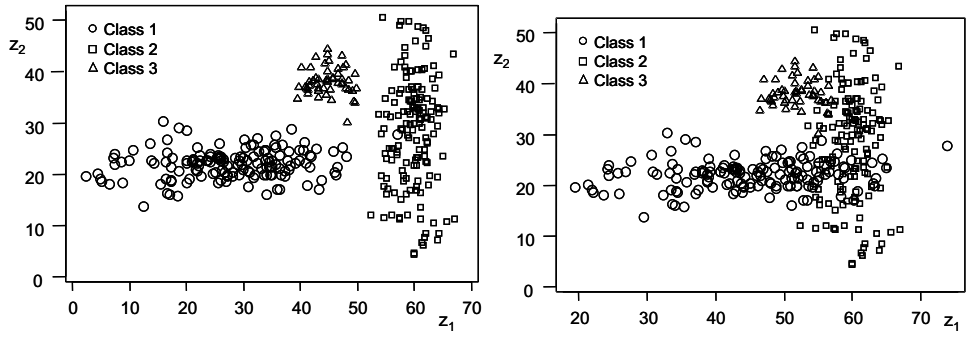

Fig. 1. Seed data sets 1 and 2 showing, respectively, well-separated and overlapping classes

In order to build interval data sets from data sets 1 and 2, each point (z_1, z_2) of these data sets is considered as the 'seed' of a rectangle. Each rectangle is therefore a vector of two intervals defined by:

$$([z_1 - \gamma_1/2, z_1 + \gamma_1/2], [z_2 - \gamma_2/2, z_2 + \gamma_2/2]) \tag{16}$$

The parameters γ_1 and γ_2 are the width and the height of the rectangle. They are drawn randomly within a given range of values. For example, the width and the height of all the rectangles can be drawn randomly within the interval $[1, 8]$. Figure 2 shows two synthetic interval data sets built from data set 1 and data set 2 when γ_1 and γ_2 are drawn randomly from $[1, 8]$.

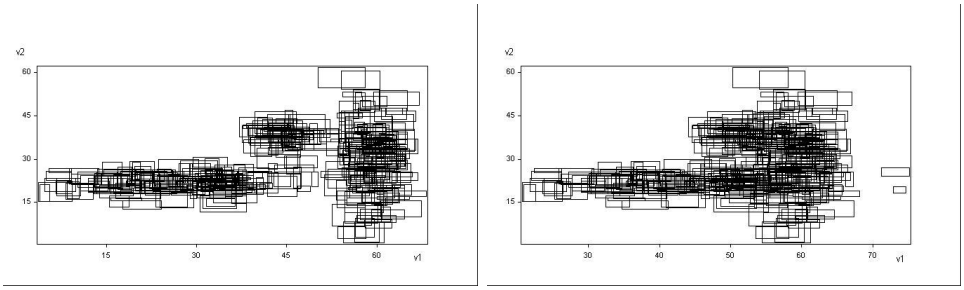

Fig. 2. Interval data sets 1 and 2, showing, respectively, well-separated and overlapping classes

In the framework of a Monte Carlo experiment, 100 replications of the previous process were carried out for parameters γ_1 and γ_2, drawn randomly 100 times from each of the following intervals: [1,8],[1,16], [1,24], [1,32], [1,40]. This process has also been repeated for seeds taken from data set 1 and data set 2.

Dynamic clustering algorithms considering different distances between vectors of intervals have been performed on these data sets. The 3-cluster partitions obtained with these clustering methods were compared with the 3-cluster partition known a priori. The comparison index used is the corrected Rand index CR given in equation (15). For each 100 replications, the average corrected Rand index CR is calculated.

Table 2 gives the values of the average CR index obtained with adaptive and non-adaptive distances for interval data sets 1 and 2 as well as γ_1 and γ_2 drawn from $[1, 8], [1, 16], [1, 24], [1, 32], [1, 40]$. As expected, in each case the average CR indices are better with adaptive distances.

Concerning the data configurations presenting well separated classes, the Hausdorff (non-adaptive) distance shows better CR indices than city-block (non-adaptive) distance regardless the range of the predefined intervals in Table 2. Moreover, the Hausdorff distance is also the best option for data configuration presenting overlapping classes as long as the widest intervals are considered.

For both type of data configurations (well separated classes and overlapping classes), the average CR indices provided by the adaptive Hausdorff distance are again better than those provided by the adaptive city-block distance for those data configurations where the range of the predefined intervals are the widest. Table 3 gives the corresponding values of the standard deviation for the average CR index.

The evaluation of the performance of the dynamic clustering methods for these different distances between vectors of intervals is achieved by an independent Student's t-test with a 5% level of significance. Tables 4 and 5 shows the suitable (null and alternative) hypothesis and the observed values of the test statistics following a Student's t distribution with 198 degrees of free-

Table 2
Comparison of the methods according to the average corrected Rand index

Predefined Intervals	Interval Data Set 1				Interval Data Set 2			
	Non-Adaptive Distances		Adaptive Distances		Non-Adaptive Distances		Adaptive Distances	
	L_1	Hausd.	L_1	Hausd.	L_1	Hausd.	L_1	Hausd.
$[1, 8]$	0.684	0.691	0.935	0.923	0.379	0.378	0.470	0.448
$[1, 16]$	0.664	0.706	0.931	0.931	0.375	0.374	0.432	0.434
$[1, 24]$	0.636	0.700	0.892	0.909	0.369	0.377	0.406	0.418
$[1, 32]$	0.622	0.685	0.773	0.912	0.361	0.385	0.389	0.412
$[1, 40]$	0.618	0.702	0.701	0.886	0.348	0.378	0.373	0.393

Table 3
Comparison of the methods according to the standard deviation of the corrected Rand index

Predefined Intervals	Interval Data Set 1				Interval Data Set 2			
	Non-Adaptive Distances		Adaptive Distances		Non-Adaptive Distances		Adaptive Distances	
	L_1	Hausd.	L_1	Hausd.	L_1	Hausd.	L_1	Hausd.
$[1, 8]$	0.0127	0.0150	0.0005	0.0010	0.0013	0.0013	0.0050	0.0019
$[1, 16]$	0.0118	0.0184	0.0009	0.0007	0.0010	0.0012	0.0014	0.0023
$[1, 24]$	0.0041	0.0154	0.0058	0.0011	0.0014	0.0010	0.0015	0.0019
$[1, 32]$	0.0019	0.0131	0.0114	0.0011	0.0013	0.0014	0.0013	0.0017
$[1, 40]$	0.0014	0.0133	0.0073	0.0032	0.0010	0.0016	0.0024	0.0024

dom for interval data sets 1 and 2, respectively. In this table, μ_1, μ_2, μ_3 and μ_4 are, respectively, the average of the CR index for the dynamic clustering algorithm considering the Hausdorff distance, city-block distance, adaptive Hausdorff distance and city-block (one componente) adaptive distance, respectvely. These tables show that, in 75% of the data simulation configurations considered in this work, the dynamic clustering algorithm based on Hausdorff distance outperforms the version of this algorithm which uses the city-block distance considering both adaptive and non-adaptive cases.

Table 4
Interval Data Set 1 (well-separated classes): statistics of Independent Student's t-tests comparing methods

Predefined Intervals	Interval Data Set 1			
	Non-Adaptive Distances $H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$	Decision	Adaptive Distances $H_0 : \mu_3 \leq \mu_4$ $H_1 : \mu_3 > \mu_4$	Decision
$[1, 8]$	3.76	Reject H_0	-44.62	Accept H_0
$[1, 16]$	19.49	Reject H_0	0.75	Accept H_0
$[1, 24]$	40.23	Reject H_0	28.99	Reject H_0
$[1, 32]$	47.68	Reject H_0	121.03	Reject H_0
$[1, 40]$	62.92	Reject H_0	232.88	Reject H_0

Table 5
Interval Data Set 2 (overlapping classes): Statistics of Independent Student's t-tests comparing methods

Predefined Intervals	Interval Data Set 2			
	Non-Adaptive Distances $H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$	Decision	Adaptive Distances $H_0 : \mu_3 \leq \mu_4$ $H_1 : \mu_3 > \mu_4$	Decision
$[1, 8]$	-2.80	Accept H_0	-40.19	Accept H_0
$[1, 16]$	-8.36	Accept H_0	6.69	Reject H_0
$[1, 24]$	42.65	Reject H_0	50.97	Reject H_0
$[1, 32]$	125.80	Reject H_0	105.25	Reject H_0
$[1, 40]$	158.54	Reject H_0	60.39	Reject H_0

4.2 Car data set

The car interval data set consists of a set of 33 car models described by 8 interval, 2 categorical multi-valued and one nominal variables (see Table 6). In this application, the 8 interval variables - *Price, Engine Capacity, Top Speed, Acceleration, Step, Length, Width* and *Height* - have been considered for clustering purposes, the nominal variable *Car Category* has been used as a *a priori* classification.

Table 6
'Car' data set with 8 interval and one nominal variables

	Price	Engine Capacity	...	Height	Category
Alfa 145	[27806, 33596]	[1370, 1910]	...	[143, 143]	Utilitarian
Alfa 156	[41593, 62291]	[1598, 2492]	...	[142, 142]	Berlina
...	...	...	...	...	...
Porsche 25	[147704, 246412]	[3387, 3600]	...	[130, 131]	Sporting
Rover 25	[21492, 33042]	[1119, 1994]	...	[142, 142]	Utilitarian
Passat	[39676, 63455]	[1595, 2496]	...	[146, 146]	Luxury

Dynamic clustering algorithms considering different distances between vectors of intervals have been performed on this data set. The 4-cluster partitions obtained with these clustering methods were compared with the 4-cluster partition known *a priori*. The comparison index used is the corrected Rand index CR given in equation (15). The *a priori* classification, indicated by the suffix attached to the car model denomination, is as follows:

`Utilitarian:`
1-Alfa 145/U 5-Audi A3/U 12-Punto/U 13-Fiesta/U 17-Lancia Y/U
24-Nissan Micra/U 25-Corsa/U 28-Twingo/U 29-Rover 25/U 31-Skoda Fabia/U

`Berlina:`
2-Alfa 156/B 6-Audi A6/B 8-BMW serie 3/B 14-Focus/B
21-Mercedes Classe C/B 26-Vectra/B 30-Rover 75/B 32-Skoda Octavia/B

`Sporting:`
4-Aston Martin/S 11-Ferrari/S 15-Honda NSK/S 16-Lamborghini/S
19-Maserati GT/S 20-Mercedes SL/S 27-Porsche/S

`Luxury:`
3-Alfa 166/L 7-Audi A8/L 9-BMW serie 5/L 10-BMW serie 7/L
18-Lancia K/L 22-Mercedes Classe E/L 23-Mercedes Classe S/L 33-Passat/L

Each clustering method is run (until the convergence to a stationary value of the adequacy criterion) 100 times and the best result, according to the adequacy criterion, is selected. The corrected Rand index CR is calculated for the best result. Table 7 shows the clusters (individual labels) given by the non-adaptive (L_1 and Hausdorff) and adaptive (one component L_1 and Hausdorff) methods. The CR indices obtained from the results displayed in Table 7 are 0.35 and 0.38 for the non-adaptive L_1 and Hausdorff methods, respectively, and 0.56 for the adaptive (one component L_1 and Hausdorff) methods. Notice that, for this data set, the dynamic clustering algorithm with non-adaptive Hausdorff distance outperforms this same algorithm with non-adaptive L_1 distance. Moreover, for the case of adaptive distances, it is furnished the same partition by the dynamic clustering algorithm (the adaptive Hausdorff and one component city-block distances presented the same performance). However, for this data set, the version of the dynamic clustering algorithm with

adaptive distances outperforms the version of this algorithm with non-adaptive distances.

Table 7
Clustering Results for the Car data set

Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4
L_1 (non-adaptive)	2/B 3/L 5/U 6/B 8/B 18/L 21/B 30/B 33/L	7/L 9/L 10/L 15/S 19/S 20/S 27/S	1/U 12/U 13/U 14/B 17/U 24/U 25/U 26/B 28/U 29/U 31/U 32/B	4/S 11/S 16/S 22/L 23/L
Hausdorf (non-adaptive)	1/U 12/B 13/U 14/B 17/U 24/U 25/U 26/B 28/U 29/U 31/U 32/B	7/L 9/L 10/L 15/S 19/S 20/S 22/L 23/L 27/S	2/B 3/L 5/U 6/B 8/B 18/L 21/B 30/B 33/S	4/S 11/S 16/S
L_1 (adaptive)	12/U 13/U 17/U 24/U 25/U 28/U 29/U 31/U	1/U 2/B 3/L 5/U 8/B 14/B 18/L 21/B 26/B 30/B 32/B 33/L	6/B 7/L 9/L 10/L 22/L 23/L	4/S 11/S 15/S 16/S 19/S 20/S 27/S
Hausdorf (adaptive)	1/U 2/B 3/L 5/U 8/B 14/B 18/L 21/B 26/B 30/B 32/B 33/L	12/U 13/U 17/U 24/U 25/U 28/U 29/U 31/U	4/S 11/S 15/S 16/S 19/S 20/S 27/S	6/B 7/L 9/L 10/L 22/L 23/L

4.3 Ecotoxicology data set

Several studies realized in French Guyana indicated abnormal levels of mercury contamination in some Amerindian populations. This contamination was connected to their high consumption of contaminated freshwater fish (Boudou and Ribeyre (1998)). In order to get a better knowledge of this phenomenon, a data set has been collected by researchers from the LEESA (Laboratoire d'Ecophysiologie et d'Ecotoxicologie des Systèmes Aquatiques) laboratory.

This data set concerns 12 species of freshwater fish, each species being described by 13 interval variables. These species are grouped into four a priori clusters of unequal sizes according to diet: two clusters (Carnivorous and Detritivorous) of size 4 and two clusters of size 2 (Omnivorous and Herbivorous). Table 6 shows part of the freshwater fish data set.

Dynamic clustering algorithms considering different distances between vectors of intervals have also been performed on this data set. The 4-cluster partitions obtained with these clustering methods were compared with the 4-cluster partition known *a priori*. Again, the comparison index used is the corrected Rand index CR given in equation (15). The *a priori* classification, indicated by the suffix attached to the freshwater specie denomination, is as follows:

Table 8
Freshwater fish data set described by 13 interval variables

Interval Variables	Individuals/Labels			
	Ageneiosusbrevifili	Cynodongibbus	...	Myleusrubripinis
Length	[22.5 : 35.5]	[19 : 32]	...	[12.3 : 18]
Weight	[170 : 625]	[77 : 359]	...	[80 : 275]
Muscle	[1425 : 5043]	[2393 : 8737]	...	[8 : 35]
Intestine	[333 : 2980.06]	[0 : 2653]	...	[0 : 0]
Stomach	[0 : 1761.1]	[478.34 : 10860.7]	...	[10.76 : 41.93]
Gills	[393.71 : 853.1]	[354.22 : 1976.38]	...	[0 : 9.45]
Liver	[642 : 7105.77]	[2684.83 : 43014]	...	[190.12 : 394.52]
Kidneys	[0 : 3969.05]	[1437.82 : 27514.6]	...	[72.3 : 112.54]
Liver/Muscle	[0.45 : 1.41]	[1.12 : 4.92]	...	[7.12 : 30.35]
Kidneys/Muscle	[0 : 2.02]	[0.6 : 3.24]	...	[2.42 : 10.23]
Gills/Muscle	[0.15 : 0.3]	[0.15 : 0.24]	...	[0 : 0.85]
Intestine/Muscle	[0.23 : 0.63]	[0 : 0.5]	...	[0 : 0]
Stomach/Muscle	[0 : 0.55]	[0.2 : 1.24]	...	[0.31 : 4.33]

`Carnivorous:`
1-Ageneiosusbrevifili/C 2-Cynodongibbus/C 3-Hopliasaïmara/C 4-Potamotrygonhystrix/C

`Detritivorous:`
7-Dorasmicropoeus/D 8-Platydorascostatus/D 9-Pseudoancistrusbarbatus/D
10-Semaprochilodusvari/D

`Omnivorous:`
5-Leporinusfasciatus/O 6-Leporinusfrederici/O

`Herbivorous:`
11-Acnodonoligacanthus/H 12-Myleusrubripinis/H

Each clustering method is run (until the convergence to a stationary value of the adequacy criterion) 100 times and the best result, according to the adequacy criterion, is selected. The corrected Rand index CR is calculated for the best result. Table 9 shows the clusters (individual labels) given by the non-adaptive (L_1 and Hausdorff) and adaptive (one component L_1 and Hausdorff) methods. The CR indices obtained from the results displayed in Table 9 are 0.488 and 0.138 for the adaptive and non-adaptive distances, respectively. Notice that, for this data set, regardless the adaptive (or the non-adaptive) distances used, the dynamic clustering algorithm furnishes the same partition (the Hausdorff and city-block distances presented the same

performance). However, for this data sets, the version of the dynamic clustering algorithm with adaptive distances outperforms the version of this algorithm with non-adaptive distances.

Table 9
Clustering Results for the Ecotoxicology data set

Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4
L_1 (non-adaptive)	1/C 4/C 7/D 8/D 10/D	2/C	3/C	5/O 6/O 9/D 11/H 12/H
Hausdorf (non-adaptive)	1/C 4/C 7/D 8/D 10/D	2/C	3/C	5/O 6/O 9/D 11/H 12/H
L_1 (adaptive)	5/O 6/O	9/D 11/H 12/H	1/C 2/C 3/C	4/C 7/D 8/D 10/D
Hausdorf (adaptive)	5/O 6/O	9/D 11/H 12/H	1/C 2/C 3/C	4/C 7/D 8/D 10/D

For the case of the dynamic clustering algorithm considering the Hausdorff adaptive distances performed on the Ecotoxicology interval data set, Tables 10 and 11, respectively, give the prototype descriptions and the corresponding weight vectors of the (Hausdorff) adaptive distances associated to each class, according to the 13 interval variables.

Table 10
Description of each class prototype according to the 13 interval variables

Interval Variables	Prototype description			
	Class 1	Class 2	Class 3	Class 4
Length	[20.9 : 24.75]	[22.5 : 35.5]	[12.05 : 18.25]	[19.27 : 30.82]
Weight	[207.5 : 311.5]	[170 : 625]	[55 : 210]	[229 : 602.5]
Muscle	[743 : 1549]	[1269.65 : 5937]	[57 : 84]	[490.5 : 929.73]
Intestine	[22.5 : 130.6]	[2.97 : 2650.03]	[0 : 95]	[71.59 : 927.22]
Stomach	[112.25 : 434.13]	[0 : 1761.1]	[55.87 : 127.67]	[0 : 579.80]
Gills	[10 : 93.24]	[346.79 : 900.02]	[0 : 9.45]	[55.09 : 133.98]
Liver	[352.81 : 871.03]	[642 : 7105.77]	[101.26 : 807]	[1744.57 : 8752.98]
Kidneys	[538.08 : 1870.74]	[1071 : 22015]	[71.25 : 113.59]	[1072.46 : 7709.51]
Liver/Muscle	[0.41 : 0.76]	[0.65 : 2.45]	[1.81 : 18.82]	[1.87 : 9.82]
Kidneys/Muscle	[0.41 : 0.76]	[0.60 : 3.24]	[2.42 : 10.23]	[1.08 : 11.64]
Gills/Muscle	[0.02 : 0.07]	[0.15 : 0.24]	[0 : 0.85]	[0.11 : 0.19]
Intestine/Muscle	[0.09 : 0.12]	[0.10 : 0.50]	[0 : 2.16]	[0.17 : 1.44]
Stomach/Muscle	[0.12 : 0.37]	[0 : 0.55]	[0.31 : 4.33]	[0 : 0.7]

In Noirhomme-Fraiture (2002), visualization techniques for interval data are

Table 11
Ecotoxicology data set: vectors of weights $\boldsymbol{\lambda}_k$ of adaptive distance d_k $(k = 1, \ldots, 4)$ according to the 13 interval variables

Interval Variables	Vectors of weights			
	$\boldsymbol{\lambda}_1$	$\boldsymbol{\lambda}_2$	$\boldsymbol{\lambda}_3$	$\boldsymbol{\lambda}_4$
Length	3.911076	3.784355	9.578303	5.855540
Weight	0.099555	0.022819	0.362272	0.022301
Muscle	0.016830	0.031758	0.375701	0.077517
Intestine	0.062889	0.146273	0.200882	0.145701
Stomach	0.042986	0.012046	0.319487	0.108850
Gills	0.158115	0.097757	0.295046	0.468606
Liver	0.032341	0.003193	0.032691	0.006864
Kidneys	0.008428	0.004982	0.082949	0.007686
Liver/Muscle	41.066313	33.423074	3.175321	2.804027
Kidneys/Muscle	28.321594	39.767796	2.033658	8.990483
Gills/Muscle	410.663114	1466.437071	12.970620	374.103973
Intestine/Muscle	68.443852	488.812463	18.866356	184.489640
Stomach/Muscle	40.064695	139.660709	9.453641	144.814440

presented, especially a type of graphic called Zoom Star. In this graphical representation, each axis corresponds to an interval variable. In each axis, the lower and upper bounds of the interval value assumed by an interval variable for a given object are represented. The lower bounds (as well the upper bounds) of the intervals assumed by each interval variable are linked to form a polygon. The Zoom Star shows the area between the upper-bound and lower-bound polygons. Figure 5 gives the visualization of the prototype of each cluster from Table 10 according to the Zoom Star method.

All the interval variables for the prototype of Cluster 1 show intervals with low spread, whereas they show intervals with medium spread for the prototype of Cluster 4. For the prototype of Cluster 2, most of the interval variables that do not represent ratios show a high spread, whereas the interval variables that express ration show a low spread. Concerning the prototype of Cluster 3, the role of the ratio and non-ratio interval variables are inverted in comparison to their role in the prototype of Cluster 2.

In conclusion, for this data set, the performance of the adaptive methods measured by the CR index is superior to the non-adaptive methods.

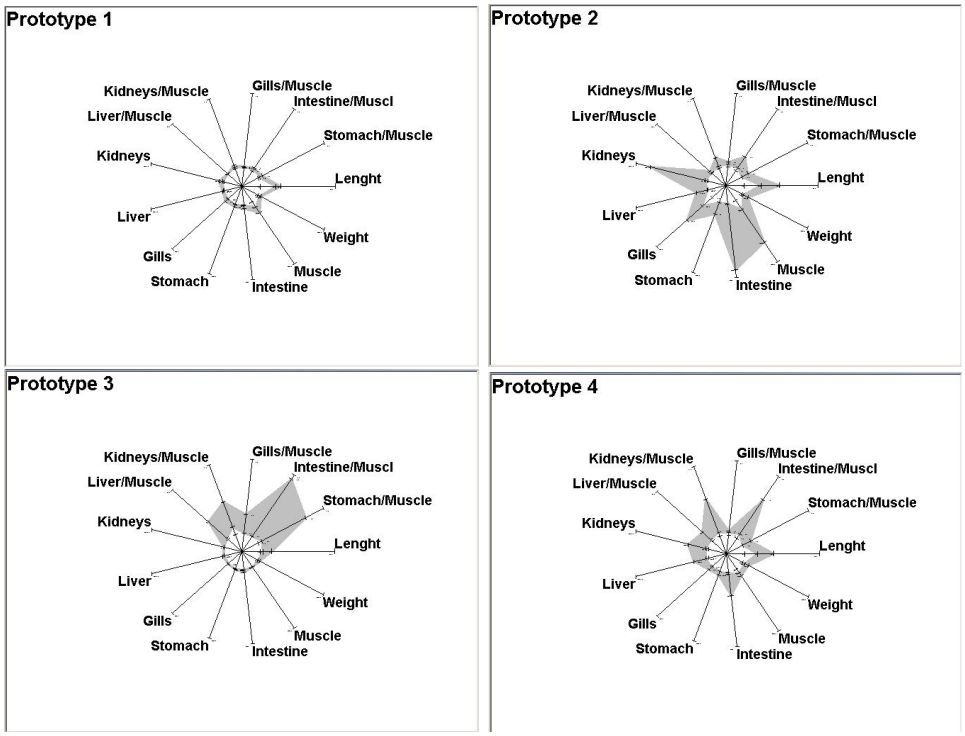

Fig. 3. Ecotoxicology data set: description of prototypes for each cluster according to Zoom Star method

5 Concluding remarks

In this paper, a clustering method for interval data using a dynamic clustering algorithm with adaptive Hausdorff distances was presented. The algorithm locally optimizes an adequacy criterion that measures the fitting between the classes and their representatives (prototypes). To compare classes and prototypes, adaptive distances based on a weighted version of the Hausdorff distance for interval data are introduced.

The dynamic clustering algorithm with adaptive Hausdorff distances starts from an initial partition and alternates a representation step and an allocation step until convergence when the adequacy criterion reaches a stationary value representing a local minimum. The representation step has two stages. In the first stage, the partition and the Hausdorff distances are fixed and the algorithm looks for the best prototype of each class which minimizes the adequacy criterion. The solution for the best prototype of each class, presented in this paper, is a vector of intervals whose lower bounds, for a given variable, are the difference between the median of midpoints of the intervals computed for the objects belonging to this class and the median of their half-lengths, and whose upper bounds, for a given variable, are the sum of the median of midpoints of the intervals computed for the objects belonging to this class plus the median of their half-lengths. In the second stage, the partition and the

prototype of each class are fixed and the algorithm looks for the best Hausdorff distance associated to each class which minimizes the adequacy criterion. The Hausdorff distance associated to each class is parameterized by a vector of weights and the best solution for this vector of weights provided by the clustering method is also presented in this paper. In the allocation step, the individuals are assigned to the classes according to their (minimum) adaptive Hausdorff distance to the prototypes.

Experiments with real and synthetic interval data sets showed the usefulness of this clustering method. The accuracy of the results furnished by the dynamic clustering algorithm based on adaptive Hausdorff distance is assessed by the CR index and compared with the results provided by this algorithm considering non-adaptive Hausdoff distance and adaptive and non-adaptive city-block distances.

Concerning the synthetic interval data sets, the CR index is calculated in the framework of a Monte Carlo experiment with 100 replications. For the data configurations showing well separated classes, the Hausdorff distance outperforms the city-block distance for the non-adaptive version of the dynamic clustering algorithm. Moreover, for the non-adaptive version of the dynamic clustering algorithm, the Hausdorff distance is also the best option for data configurations presenting overlapping classes as long as the widest intervals are considered. For both types of data configurations (well separated classes and overlapping classes) the adaptive Hausdorff distance outperforms the adaptive city-block distance also as long as the widest intervals are considered.

Concerning the car interval data set, the Hausdorff distance outperforms the city-block distance for the non-adaptive version of the dynamic clustering algorithm. Moreover, these distances presented the same performance when the adaptive version of the dynamic clustering algorithm is applied on this data set. Concerning the ecotoxicology data set, the Hausdorff and city-block distances presented the same performance for the non-adaptive and adaptive version of the dynamic clustering algorithm. However, for both data sets, the version of the dynamic clustering algorithm with adaptive distances outperforms the version of this algorithm with non-adaptive distances.

References

Bobou, A. and Ribeyre, F. 1998. Mercury in the food web: accumulation and transfer mechanisms, in Sigrel A. and Sigrel H. (Eds), Metal Ions in Biological Systems. M. Dekker, New York, 289-319.

Billard, L. and Diday, E. 2003. From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. Journal of American Statistical Association, **98**, (462), 470–487

Bock, H. H. 1993. Classification and Clustering: Problems for the Future. In: Diday et al Eds., New Approaches in Classification and Data Analysis, Springer-Verlag, 3-24.

Bock, H. H. and Diday, E. 2000. Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data. Springer-Verlag, Heidelberg.

Bock H.-H. 2002. Clustering algorithms and kohonen maps for symbolic data. Proc. ICNCB, Osaka, 203-215. J. Jpn. Soc. Comp. Statistic, **15**, 1–13

Chavent, M. and Lechevallier, Y. 2002. Dynamical Clustering Algorithm of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance. In: Sokolowsky and H.H. Bock Eds., Classification, Clustering and Data Analysis. K. Jaguja, A. Springer-Verlag, Heidelberg, 53-59.

De Carvalho, F. A. T. 1994. Proximity coefficients between Boolean symbolic objects. In: E. Diday et al Eds., New Approaches in Classification and Data Analysis. Heildeberg, Springer-Verlag, 387-394.

De Carvalho, F. A. T. 1998. Statistical proximity functions of boolean symbolic objects based on histograms. In: Rizzi, A. et al Eds., New Andvances in Data Science and Classification. Heildelberg. Springer-Verlag, 391 - 396.

De Carvalho, F. A. T. and Souza, R. M. C. R. 1998. New metrics for Constrained Boolean Symbolic Objects. In: Studies and Reserach: Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98), Office for Official Publications of the European Communities, Luxemburg, 175-187.

Diday, E. 1971. La méthode des Nuées dynamiques *Revue de Statistique Appliquée, **19**, 2, 19–34.*

Diday, E. and Govaert, G. 1974. Classification avec Distances Adaptatives. Compte redu de l'Académie des Sciences, Paris, tome 278, série A, p. 993.

Diday, E. and Simon, J. J. 1976. Clustering Analysis. In: Fu, K. S. Eds. Digital Pattern Recognition. Springer-Verlag, Heidelberg, 47-94.

Diday, E. and Govaert, G. 1977. Classification Automatique avec Distances Adaptatives. R.A.I.R.O. Informatique Computer Science, **11** (4), 329-349.

Diday, E. 1988. The symbolic approach in clustering and related methods of data analysis. In: H. H. Bock Ed., Classification methods of Data Analysis. North Holland, Amsterdam, 673-684.

Diday, E. and Brito, M. P. 1989. Symbolic Cluster Analysis. In: O. Opitz Eds., Conceptual and Numerical Analysis of Data. Springer-Verlag, Heidelberg, 45-84

El-Sonbaty, Y. and Ismail, M. A. 1998. Fuzzy Clustering for Symbolic Data. IEEE Transactions on Fuzzy Systems, **6**, 195-204.

Everitt, B. 2001. Cluster Analysis. Halsted, New York.

Gordon, A. D. 1999. Classification. Chapman and Hall/CRC, Boca Raton, Florida.

Gordon, A. D. 2000. An Iteractive Relocation Algorithm for Classifying Symbolic Data In: W. Gaul et al Eds., Data Analysis: Scientific Modeling and Practical Application. W. Gaul, Springer-Verlag, Berlin, 17-23.

Govaert, G. 1975. Classification automatique et distances adaptatives. Thèse de 3ème cycle, Mathématique appliquée, Université Paris VI.

Gowda, K. C. and Diday, E. 1991. Symbolic clustering using a new dissimilarity measure. Pattern Recognition, **24**, (6), 567-578.

Gowda, K. C. and Diday, E. 1992. Symbolic clustering using a new similarity measure. IEEE Transactions on Systems, Man and Cybernetics, **22**, 368-378.

Hubert, L. and Arabie. P. 1985. Comparing Partitions. Journal of Classification, **2**, 193-218.

Huttenlocher, D.P., Klanderman G.A. and Rucklidge W.J. 1993. Comparing images using the Hausdorff Distance. IEE Transaction on Pattern Analysis and Machine Intelligence, **15**, 850–863

Ichino, M. and Yaguchi, H. Generalized Minkowski metrics for mixed feature type data analysis. IEEE Transactions on Systems Man and Cybernetics, **24**, (4), 698-708.

Jain, A.K., Murty, M.N. and Flynn, P.J. 1999. Data Clustering: A Review. ACM Computing Surveys, **31**, (3), 264-323.

Nadler S.B. Jr. 1978. Hyperspaces of sets. Marcel Dekker, Inc., New York

Noirhomme-Fraiture, M. 2002. Visualization of Large Data Sets: The Zoom Star Solution. Electronic Journal of Symbolic Data Analysis, **0**, 26-39.

Rote G., 1991. Computing the minimum Hausdorff distance between two point sets on a line under translation. Information Processing Letters, **38**, 123–127

Souza, R.M.C.R. and De Carvalho, F. A. T.: Clustering of interval data based on city-block distances. Pattern Recognition Letters, **25** (3) (2004) 353–365

Spaeth, H. 1980. Cluster analysis algorithms. Wiley, New York.

Verde, R., De Carvalho, F. A. T. and Lechevallier, Y. 2001. A Dynamical Clustering Algorithm for symbolic data. Tutorial on Symbolic Data Analisys, GfKl Conference, Munich.